DISSENSUS AI WORKING PAPER SERIES

DAI-2505

Context-Dependent Affordance Computation in Vision-Language Models

Murad Farzulla[1,2,*]

[1]Dissensus AI, London, UK [2]King's College London, London, UK

[*]Correspondence: murad@dissensus.ai ORCID: 0009-0002-7164-8704

January 2026

Abstract

We characterize the phenomenon of *context-dependent affordance computation* in vision-language models (VLMs). Through a large-scale computational study ($n = 3{,}213$ scene-context pairs from COCO-2017) using Qwen-VL 30B and LLaVA-1.5-13B subject to systematic context priming across 7 agentic personas, we demonstrate massive affordance drift: mean Jaccard similarity between context conditions is 0.095 (95% CI: [0.093, 0.096], $p < 0.0001$), indicating that $> 90\%$ of lexical scene description is context-dependent. Sentence-level cosine similarity confirms substantial drift at the semantic level (mean $= 0.415$, 58.5% context-dependent). Stochastic baseline experiments (2,384 inference runs across 4 temperatures and 5 seeds) confirm this drift reflects genuine context effects rather than generation noise: within-prime variance is substantially lower than cross-prime variance across all conditions. Tucker decomposition with bootstrap stability analysis ($n = 1{,}000$ resamples) reveals stable orthogonal latent factors: a "Culinary Manifold" isolated to chef contexts and an "Access Axis" spanning child-mobility contrasts. These findings establish that VLMs compute affordances in a substantially context-dependent manner—with the difference between lexical (90%) and semantic (58.5%) measures reflecting that surface vocabulary changes more than underlying meaning under context shifts—and suggest a direction for robotics research: dynamic, query-dependent ontological projection (JIT Ontology) rather than static world modeling. We do not claim to establish processing order or architectural primacy; such claims require internal representational analysis beyond output behavior.

Keywords: Vision-language models, Affordances, Context-dependent processing, Scene understanding, Functional semantics, Robotics

Acknowledgements

The author thanks Claude (Anthropic) for assistance with analytical framework development, tensor decomposition analysis, and technical writing. The author also thanks the Visual Genome project for providing human affordance annotations for baseline comparison. This paper is part of the Adversarial Systems Research program at Dissensus AI and the Adversarial Systems & Complexity Research Initiative (ASCRI). All errors, omissions, and interpretive limitations remain the author's responsibility.

Data & Code Availability. Analysis code and data are available at `https://github.com/studiofarzulla/semantic-vision`.

1 Introduction

Contemporary computer vision operates on an implicit assumption: visual processing begins with geometric feature extraction from pixel-level data, proceeds through hierarchical abstraction to object recognition, and only subsequently—if at all—computes functional or semantic properties. This pipeline reflects a Cartesian conception of space as a neutral container:

$$\mathscr{P}_{\text{std}} : I \to F_{\text{pixel}} \to F_{\text{feature}} \to O_{\text{object}} \to C_{\text{context}} \to A_{\text{affordance}} \tag{1}$$

This ordering is not theoretically neutral. It embeds assumptions about perception that have been challenged by ecological psychology (Gibson, 1979), phenomenology (Heidegger, 1927; Merleau-Ponty, 1945), and cognitive neuroscience (Goodale and Milner, 1992). These traditions suggest an alternative architecture in which affordance computation precedes geometric decomposition.

We investigate whether this alternative architecture manifests in vision-language models (VLMs). Our **Research Question**: Do VLMs exhibit context-dependent affordance computation consistent with a semantic-first architecture, where functional interpretation precedes and structures geometric representation?

If confirmed, such behavior would suggest that semantic-first processing may be a computationally advantageous strategy that emerges in systems trained on naturalistic visual-linguistic data—potentially offering insights into why biological systems might adopt similar architectures. The implied processing order would be:

$$\mathscr{P}_{\text{SFS}} : I \to T_{\text{token}} \to C_{\text{context}} \to G_{\text{geo}|C} \to A_{\text{aff}|C,\Theta} \to S_{\text{spatial}|A} \tag{2}$$

where the conditioning notation $X_{a|b}$ denotes that representation a is computed conditional on prior establishment of b, and Θ represents agent goal states.

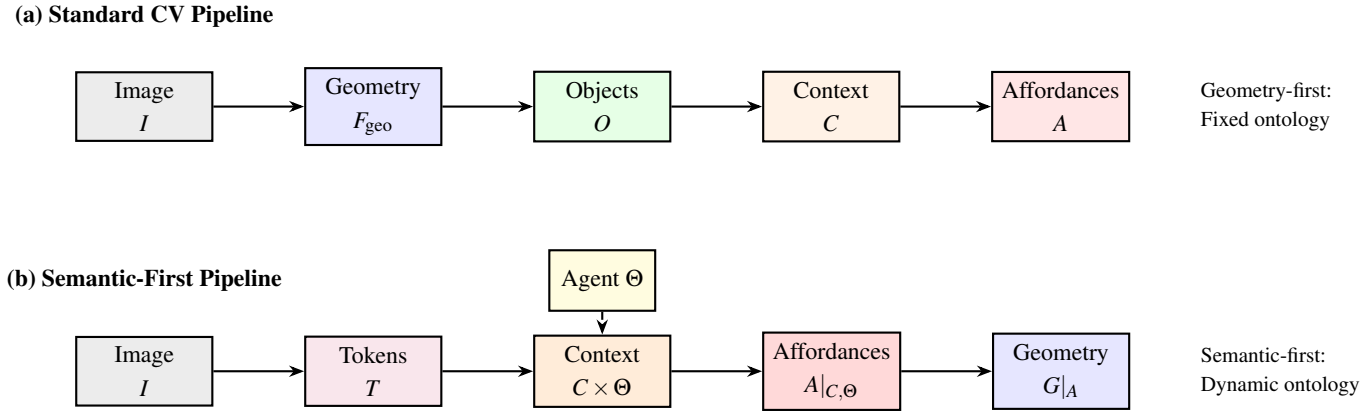

Figure 1: Comparison of visual processing pipelines. (a) Standard computer vision computes geometry before semantics, producing a fixed scene ontology. (b) The proposed Semantic-First architecture conditions geometric processing on agent context Θ, enabling dynamic, task-relevant representations. Arrow direction indicates computational dependency.

Figure 1 illustrates the architectural difference between the standard geometry-first pipeline and our proposed semantic-first alternative. Notably, Gokhale (2024) independently arrives at the same terminological and conceptual framing from the recognition side of computer vision, arguing that "meaning, and by proxy, natural language, serves as a critical source of knowledge for modern CV" and characterizing the field's trajectory as a paradigm shift from recognition to reasoning. The convergence

is telling: Gokhale's finding that language-guided depth estimators fail specifically with low-level geometric descriptions—while succeeding with semantically rich ones—provides independent evidence that semantic processing precedes and structures geometric decomposition in modern VLMs, precisely the architectural claim we formalize here. The contributions of this paper are: (1) **empirical demonstration** that VLMs exhibit massive context-dependent affordance drift, with $> 90\%$ of functional scene ontology varying by agent context; (2) **theoretical proposal** of semantic-first processing as a candidate model for biological spatial cognition, motivated by convergent evidence from ecological psychology and phenomenology; and (3) **a speculative direction** for robotics and AI systems via Just-In-Time (JIT) Ontology.

2 Theoretical Framework

2.1 Formal Definitions

Definition 2.1 (Visual Field)**.** A visual field $\mathscr{V}$ is the totality of visual information available to an agent at time t, represented as image $I \in \mathbb{R}^{H\times W\times C}$.

Definition 2.2 (Agent State)**.** An agent state $\Theta = (\theta_{\text{goal}}, \theta_{\text{motor}}, \theta_{\text{history}})$ comprises current goal structure, available motor repertoire, and relevant experiential history.

Definition 2.3 (Affordance Mapping)**.** An affordance function $\alpha : G \times C \times \Theta \to \mathscr{A}$ maps geometric primitives, context, and agent state to affordance vectors encoding primary action possibility, alternative actions, and required motor engagement.

Definition 2.4 (Action-Distance)**.** The action-distance $D_{\text{action}}(g_i, g_j \,|\, \Theta)$ between objects g_i and g_j is the minimum action sequence length required to bring g_i into interaction with g_j:

$$D_{\text{action}}(g_i, g_j \,|\, \Theta) = \min_{\pi\in\Pi} |\pi| \tag{3}$$

where Π is the set of feasible action sequences.

Action-distance violates Euclidean metric properties: it exhibits asymmetry ($D_{\text{action}}(g_i, g_j) \neq D_{\text{action}}(g_j, g_i)$), goal-dependence, and context-dependence.

2.2 Core Hypotheses and Predictions

Our **Core Hypothesis**: If spatial awareness operates semantic-first—with functional semantics grounding geometric structure rather than the reverse—then we should observe massive context-dependence in affordance computation. We test this prediction using VLMs as model systems, while acknowledging that VLM behavior provides evidence for computational viability, not direct evidence about biological mechanisms.

The framework generates one hypothesis (H2) and three predictions (P1, P3, P4). **H2 is directly tested in this paper**; P1, P3, and P4 are theoretical predictions derived from the framework, presented here to motivate future empirical work (see Section 6.5 for detailed discussion of empirical coverage).

Hypothesis 2.1 (Context-Dependence of Geometry—H2 (Tested))**.** The same geometric primitive receives different functional encodings under different contexts:

$$\exists g, C_1, C_2 : \alpha(g, C_1, \Theta) \neq \alpha(g, C_2, \Theta) \tag{4}$$

The following predictions are generated by the Semantic-First framework but **not tested in this study**:

Prediction 2.1 (Semantic Priority—P1)**.** Removal of functional-semantic grounding causes dissolution of coherent spatial representation. Ablating context encoding produces degradation exceeding that from ablating any downstream component. *(Untested; requires ablation studies on VLM components.)*

Prediction 2.2 (Goal-Relativity of Space—P3)**.** Spatial relations are computed over affordances, not geometry. The distance metric relevant to spatial reasoning is action-distance, not Euclidean distance. *(Untested; requires spatial reasoning tasks comparing action-distance vs. Euclidean predictions.)*

Prediction 2.3 (Affordance Primacy in Attention—P4)**.** Attentional allocation is determined by affordance-relevance to current goals, not geometric salience. *(Untested; requires eye-tracking or attention mechanism analysis.)*

Scope and Claim Type

Before proceeding, we clarify the epistemic status of our claims. This paper makes *descriptive* claims about how a particular class of computational systems (vision-language models) processes visual scenes. We are not making *phenomenological* claims about the structure of conscious experience, nor *neuroscientific* claims about biological neural implementation.

When we invoke phenomenological concepts (ready-to-hand, motor intentionality) or ecological psychology (affordances, direct perception), we do so as *theoretical inspiration*, not as hypotheses under direct test. The relationship is analogical: phenomenology suggests that practical engagement structures perception; we operationalize an analogue of this suggestion and test whether VLMs exhibit corresponding behavior.

Our findings are *suggestive for robotics*: if spatial cognition benefits from context-first processing, then architectures implementing this principle may perform better on embodied tasks. But we do not claim to have proven that biological cognition works this way, only that a particular artificial system exhibits massive context-dependence that current vision pipelines do not accommodate.

This modesty is methodological, not rhetorical. The question of whether VLM behavior reveals anything about biological cognition is empirically open. What we *can* claim is that context-dependence of the magnitude we observe (90%+) is computationally significant regardless of biological analogy.

3 Related Work

3.1 Ecological Psychology and the Ontology of Affordances

Gibson's ecological approach (Gibson, 1979) provides the foundational concept of *affordance*—action possibilities the environment offers to an agent. A crucial interpretive point requires emphasis: for Gibson, affordances are *objective* properties of the agent-environment system, not subjective construals (Turvey, 1992; Stoffregen, 2003). A chair affords sitting for a human-sized agent whether or not the agent notices, intends, or is capable of sitting at that moment. Affordances exist as relational properties—stance-independent facts about what actions the environment makes available to agents with particular bodily configurations.

Our study does not contradict Gibson's ontological claim. We do not argue that affordances *themselves* change with context—the chair's sittability remains constant. Rather, we investigate which affordances are *attended to*, *reported*, and *computationally salient* under different goal states. This is a question about attentional modulation and salience filtering, not affordance ontology. Chemero (2003) distinguishes affordances-as-properties from affordances-as-perceived, and our empirical work concerns the latter: how cognitive systems select from the space of available affordances when processing a scene.

Cisek's affordance competition hypothesis (Cisek, 2007) provides the computational framework most relevant to our findings. Cisek argues that the brain simultaneously specifies multiple potential actions (affordances) and selects among them through biased competition. On this account, visual scenes present a *field* of competing affordances, with attentional and goal-related signals modulating which affordances dominate processing. Our 90% drift finding quantifies this competition: shifting context does not create or destroy affordances but radically reshapes which affordances win the competition for representational resources.

Earlier work (Gibson, 1966) developed perceptual systems as active, exploratory processes. This supports our emphasis on goal-directed visual processing: perception is not passive reception but active interrogation of the environment for action-relevant structure.

3.2 Embodied and Enactive Cognition

The enactive approach (Varela et al., 1991; Thompson, 2007) posits that cognition brings forth a world through sensorimotor coupling rather than representing a pre-given environment. While we do not commit to strong enactivism's anti-representationalist claims, the framework's emphasis on action-perception coupling informs our hypothesis that spatial cognition is structured by affordance relations rather than geometric primitives.

O'Regan and Noë's sensorimotor contingency theory (O'Regan and Noë, 2001; Noë, 2004) establishes that perceptual content is constituted by practical mastery of sensorimotor regularities. A key implication: what we "see" is not a static picture but a structured space of possible interactions. This directly supports P3's claim that spatial relations are computed over action-possibilities rather than Euclidean geometry.

Clark's extended cognition thesis (Clark, 1997) argues that cognition is distributed across brain, body, and environment. This motivates our formal inclusion of agent state Θ—comprising goal structure, motor repertoire, and experiential history—as a first-class parameter in affordance computation. The agent-environment boundary is not fixed at the skull.

3.3 Phenomenological Inspiration

We draw inspiration from phenomenological analyses while remaining explicit about the limits of this engagement. Phenomenology offers *structural descriptions* of experience; our study offers *computational measurements* of VLM behavior. These are different kinds of claims, and we do not assert that our findings directly test phenomenological hypotheses.

Heidegger's analysis of *Zuhandenheit* (ready-to-hand) (Heidegger, 1927) distinguishes equipment encountered in use from objects contemplated theoretically. The hammer is disclosed *as* for-hammering through practical engagement, not first perceived geometrically then interpreted functionally. Importantly, ready-to-hand is not a claim about temporal processing order or neural implementation—it is a claim about the *structure of disclosure*, about how entities show up for Dasein. We do not claim our VLM experiments test Heidegger's phenomenology. Rather, Heidegger's analysis *suggests* that functional structure might be more fundamental to practical cognition than geometric structure—a suggestion we operationalize computationally and find empirically supported in a specific artificial system.

Merleau-Ponty's motor intentionality (Merleau-Ponty, 1945) describes space as structured by bodily readiness for action rather than as a neutral container. Again, this is phenomenological description, not cognitive science hypothesis. But it motivates our investigation of whether computational vision systems exhibit analogous structure: does context (analogous to bodily orientation toward tasks) reshape spatial representation?

Dreyfus's critique of classical AI (Dreyfus, 1992, 2007) argued that GOFAI failed precisely because

it assumed cognition operates on context-free symbolic representations rather than being embedded in skillful, embodied coping. Dreyfus contended that Heideggerian AI would require "making it more Heideggerian"—grounding computation in something like practical involvement rather than detached representation. Our semantic-first architecture resonates with this critique: we argue that treating spatial cognition as geometry-first representation-building misses the fundamentally context-laden, action-oriented character of biological perception.

3.4 Predictive Processing: An Alternative Account

Clark's predictive processing framework (Clark, 2013) offers an alternative theoretical lens on our findings. On predictive processing accounts, perception is hierarchical Bayesian inference: the brain generates predictions about sensory input and updates based on prediction error. Context and goals shape perception by modulating prior expectations.

From this perspective, our affordance drift could be explained as prior-shifting: different agentic contexts establish different prior probability distributions over scene contents, causing the same sensory input to yield different posterior representations. The chef-context loads priors expecting culinary affordances; the security-context loads priors expecting threats and vulnerabilities. This prior-shifting interpretation also connects to the θ_{history} component of our agent state definition: if developmental experience functions as training data for biological neural networks (Farzulla, 2025), then accumulated experiential history—including adverse experiences—would systematically shape which affordances become salient. We do not test this connection here, but note it as a direction linking affordance computation to developmental trajectories.

Recent behavioral evidence supports the claim that schema-based predictions operate at the level of general object recognition rather than fine-grained perceptual features: Suárez et al. (2026) show that contextual congruency (object–scene match) selectively enhances recognition accuracy and processing efficiency while leaving fine-grained perceptual detail retrieval unaffected. This dissociation is consistent with our proposal that context modulates affordance-level representations—the functional layer at which objects are recognized and categorized for action—rather than low-level geometric features.

We do not adjudicate between ecological and predictive processing accounts. Both are consistent with our empirical findings, and both predict context-dependent spatial representation. The key point is that *neither* framework supports the standard vision pipeline's assumption of context-independent geometric processing preceding semantic interpretation.

3.5 Cognitive Neuroscience

The two-streams hypothesis (Goodale and Milner, 1992) distinguishes dorsal (action/location) from ventral (identity) processing. Rather than claiming the dorsal stream is "really" an affordance processor, we note that the existence of parallel action-oriented and identity-oriented pathways supports the general claim that perception is not a single geometry-to-semantics pipeline.

Canonical neurons in premotor cortex (Murata et al., 1997) fire in response to graspable objects even without action intention, suggesting that affordance-relevant properties are encoded early in visual processing. Mirror neuron research (Rizzolatti and Craighero, 2004) demonstrates that action-related encoding integrates with object perception at relatively early stages. This neural evidence is consistent with affordance-sensitive processing preceding detailed geometric analysis, though we emphasize that VLM architectures need not implement mechanisms analogous to biological neural circuits.

3.6 Computer Vision Limitations

Current architectures treat affordances as late additions (Hassanin et al., 2022). Scene recognition (Zhou et al., 2017) runs parallel to, not prior to, object detection. Affordance detection (Nguyen et al., 2017) operates on object detections—the pipeline remains object-first. Scene graphs (Xu et al., 2017) encode geometric and categorical relations rather than functional and action-theoretic ones.

Recent work on task-conditioned perception begins to address these limitations, but typically treats task as an auxiliary input rather than as constitutive of representation itself. Our results suggest a more radical restructuring may be warranted: the 90% context-dependent signal might function not as an add-on but as the primary representational content.

3.7 Vision-Language Models for Robotic Affordances

Recent work in robotics has begun exploring context-sensitive affordance computation using vision-language models, addressing practical challenges that intersect with our theoretical concerns. We position our contribution as complementary to this emerging literature: where these systems *build* context-aware affordance architectures, we *study* the phenomenon of context-dependence itself and provide formalizations that could inform future system design.

VLM-Grounded Affordance Prediction. AffordanceLLM (Qian et al., 2024) leverages VLM world knowledge to ground affordance maps for in-the-wild objects, demonstrating that learned visual-linguistic representations encode functionally relevant structure. SEA (Self-Explainable Affordance) (Zhang et al., 2024) extends this by requiring robots to articulate their affordance predictions through embodied captions, bridging explainability with affordance grounding. These approaches validate our core premise: VLMs trained on naturalistic data spontaneously develop affordance-relevant representations. Our contribution complements this work by quantifying the *magnitude* of context-dependence—the 90% drift we observe provides empirical grounding for why context-sensitivity is not merely useful but essential.

Affordances as Intermediate Representations. RT-Affordance (Nasiriany et al., 2024) from Google DeepMind proposes affordances as versatile intermediate representations for robot manipulation, achieving 69% success compared to 15% for language-conditioned policies on challenging manipulation tasks. RoboPoint (Yuan et al., 2024) trains VLMs to predict keypoint affordances from language instructions, demonstrating that the same scene admits different action-relevant points depending on task specification. VoxPoser (Huang et al., 2023) takes this further by composing 3D value maps at inference time from language model affordance inferences, synthesizing dense robot trajectories zero-shot for open-set instructions—a direct instantiation of query-time ontological projection. These systems operationalize a form of Just-In-Time Ontology: rather than building comprehensive scene representations, they project task-specific affordance structures at query time. Our theoretical framework (action-distance, JIT Ontology) provides formal grounding for this architectural choice, while our empirical findings suggest it may be even more critical than current systems assume—if 90% of functional ontology is context-dependent, static world models compute primarily irrelevant structure.

Language-Affordance Grounding. The foundational work SayCan (Ahn et al., 2022) demonstrated that large language models can be grounded in robotic affordances by combining semantic knowledge ("Say"—what is useful) with affordance functions ("Can"—what is feasible). This decomposition aligns with our distinction between affordances-as-properties and affordances-as-salient: the environment presents objective action possibilities, but which possibilities become computationally active depends on task context. SayCan's 84% accuracy in skill selection versus 50% without grounding provides independent evidence for the functional importance of affordance-conditioned processing.

Context-Conditioned Affordance Ranking. Most directly relevant to our findings, Huang et al.

(2024) introduce the TAR (Task-oriented Affordance Ranking) dataset and CGR (Context-embed Group Ranking) framework, demonstrating that object affordances vary in priority across task contexts. Their dataset of 50,404 images across 25 tasks with 661k object instances provides ecological validity for context-dependent affordance computation. Critically, they show that treating objects within an affordance category as equivalent—ignoring task context—degrades performance. Our tensor decomposition results (orthogonal Culinary, Access, and Salience dimensions) complement this finding by revealing the *structure* of context-dependence: different task contexts do not merely re-weight a single affordance dimension but project scenes onto qualitatively different functional manifolds.

Our Contribution Relative to This Literature. The robotics community has developed sophisticated *systems* for context-aware affordance computation. Our work makes a distinct contribution: we *characterize* context-dependence as a phenomenon, providing:

1. **Quantification**: The 90% context-dependent signal establishes a baseline magnitude that system designers may need to accommodate. Current architectures that treat context as auxiliary input may underestimate its constitutive role.

2. **Formalization**: Action-distance and JIT Ontology provide theoretical vocabulary for discussing context-dependent spatial cognition. These concepts could inform loss functions, evaluation metrics, and architectural choices in affordance-learning systems.

3. **Decomposition**: Tucker decomposition reveals interpretable latent structure (Culinary, Access, Salience factors) suggesting that context-dependence is not diffuse but organized around functional categories. This structure could inform task taxonomies and transfer learning strategies.

The convergent finding—from both system-building (RT-Affordance, RoboPoint, VoxPoser, SayCan) and phenomenon-characterization (this work)—that context radically restructures affordance computation suggests this is not an engineering convenience but a fundamental property of how functional information is organized. Architectures that ignore context-dependence do not merely lack a feature; they operate primarily in the 10% residual space while neglecting the 90% that varies with task. Shinde et al. (2025) provide a comprehensive survey of the VLM efficiency landscape, identifying broadening modality coverage and on-device deployment as key open challenges. Their analysis of performance–memory trade-offs across compact VLM architectures contextualizes our findings within the broader engineering constraint: if context-dependence is as pervasive as our results indicate, then efficient VLM designs must accommodate dynamic, context-sensitive representations rather than relying on static scene encodings—a constraint that current efficiency-focused architectures have not fully addressed.

4 Methodology

4.1 Study Design

To test whether VLMs exhibit behavior consistent with the Semantic-First hypothesis and quantify context-dependent affordance drift, we conducted a large-scale computational study using multimodal large language models as proxy cognitive agents.

Dataset: COCO-2017 validation set (Lin et al., 2014), selecting multi-object scenes with high interaction potential. Initial corpus: 500 images.

Model: Qwen-VL-30B-Instruct (Bai et al., 2023), a high-performance vision-language model capable of detailed spatial reasoning and instruction following.

Inference Parameters: All model queries used temperature $= 0.7$ to balance affordance diversity with semantic coherence. This moderate temperature encourages exploration of the affordance space

while maintaining interpretable outputs. Lower temperatures (< 0.3) risk collapsing to stereotypical responses; higher temperatures (> 1.0) produce incoherent outputs. The selected value represents a principled trade-off, though systematic temperature ablation remains for future work (see Section 6.6).

Context Primes: For each image, the model identified critical objects and their affordances under 7 distinct agentic personas (Table 1).

Table 1: Context Prime Conditions

ID	Condition	Prime Description
P0	Neutral	Objective analysis
P1	Chef	Cooking/food preparation focus
P2	Security	Vulnerability/defense assessment
P3	Child	Play/exploration focus (4-year-old)
P4	Mobility	Obstruction/access (wheelchair user)
P5	Urgent	Immediate survival tool focus (30s emergency)
P6	Leisure	Relaxation/enjoyment, no time pressure

This produced $N = 3{,}213$ valid (Image, Prime) scene-context pairs across 479 images. Of these, 360 images produced valid affordance outputs across all seven context primes; images with JSON parsing failures, incomplete prime coverage, or malformed responses were excluded from tensor analysis to ensure balanced decomposition (see Section 4.2).

4.2 Analysis Methods

Affordance Drift: We quantified the degree to which functional scene description changes across contexts using Jaccard similarity:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{5}$$

computed at both word-level (all affordance terms) and object-level (identified objects).

Hypothesis Testing: Permutation tests (10,000 iterations) assessed whether observed Jaccard values were significantly below 0.5 (the threshold indicating more difference than overlap). Bootstrap resampling (10,000 iterations) provided confidence intervals.

Tensor Decomposition: To reveal latent functional structure, affordance text outputs were embedded using sentence-transformers (Reimers and Gurevych, 2019) (all-MiniLM-L6-v2, 384 dimensions). The resulting tensor $\mathscr{T} \in \mathbb{R}^{n_{\text{images}} \times n_{\text{primes}} \times n_{\text{embed}}}$ was decomposed via Tucker decomposition (Tucker, 1966):

$$\mathscr{T} \approx \mathscr{G} \times_1 U^{(\text{image})} \times_2 U^{(\text{context})} \times_3 U^{(\text{embed})} \tag{6}$$

The context factor matrix $U^{(\text{context})} \in \mathbb{R}^{7 \times 3}$ reveals how the 7 primes project onto latent functional dimensions.

Software and Reproducibility: All analyses were conducted in Python 3.11. Key packages: `sentence-transformers` 2.2.2 (embeddings), `tensorly` 0.8.1 (Tucker decomposition), `numpy` 1.26.4 (numerical operations), `scipy` 1.12.0 (statistical tests). VLM inference used the OpenAI-compatible API via `openai` 1.12.0. Random seeds were fixed at 42 for stratified sampling and bootstrap initialization. All

code, prompts, and analysis scripts are available at https://github.com/studiofarzulla/semantic-vision.

4.3 Affordance Extraction and Normalization

This section details the computational pipeline for extracting and normalizing affordance data from model outputs, enabling reproducibility and clarifying methodological limitations.

4.3.1 Model Output Format

The VLM was prompted to return structured JSON with keys: `objects` (a list containing objects with `id`, `name`, `affordance`, and `reasoning` fields). For example:

```
{
  "objects": [
    {"id": 1, "name": "dining table",
     "affordance": "providing a flat surface for eating",
     "reasoning": "The table is rectangular..."}
  ]
}
```

4.3.2 JSON Parsing and Error Handling

Raw model outputs frequently included markdown code fences (‘‘‘`json` ... ‘‘‘) which were stripped before JSON parsing. The parsing procedure was:

1. Remove markdown code fence delimiters
2. Strip leading/trailing whitespace
3. Parse as JSON using Python’s `json.loads()`
4. On parse failure, record entry as error and exclude from analysis

Of 3,349 attempted scene-context pairs, 136 entries (4.1%) failed JSON parsing, yielding $n = 3{,}213$ valid entries across 479 images. Parse failures were distributed uniformly across primes ($\chi^2(6) = 3.2$, $p = 0.78$), indicating no systematic bias toward particular context conditions. For Tucker decomposition, only images with complete prime coverage (all 7 primes successfully parsed) were included, yielding 360 images (75% of corpus). Excluded images did not differ systematically in COCO category distribution from included images.

4.3.3 Affordance Text Extraction

For each successfully parsed response, affordance text was constructed by concatenating:

$$\text{text}_i = \bigoplus_{o \in \text{objects}_i} (\texttt{name}_o \oplus \texttt{affordance}_o \oplus \texttt{reasoning}_o) \tag{7}$$

where $\oplus$ denotes string concatenation with space separation. The combined string was converted to lowercase. This approach captures both the object identification and the functional description, treating the full model explanation as the affordance representation.

4.3.4 Tokenization for Word-Level Jaccard

Word-level Jaccard similarity was computed using minimal preprocessing:

1. **Case normalization**: All text converted to lowercase
2. **Tokenization**: Python's `str.split()` on whitespace
3. **Set construction**: Unique tokens extracted as Python `set`

The Jaccard coefficient was then computed as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{8}$$

where A and B are token sets from two affordance texts.

Methodological Limitations. This tokenization approach is deliberately simple and carries known limitations:

- **No lemmatization**: "cooking" and "cook" are treated as distinct tokens. This inflates apparent dissimilarity when the same concept appears in different grammatical forms.
- **No stop word removal**: High-frequency function words ("the", "a", "for", "with") contribute to overlap, potentially inflating similarity estimates. However, because both affordance texts derive from similar prompt structures, stop word overlap is approximately balanced across conditions.
- **No stemming**: Morphological variants ("obstruct"/"obstruction"/"obstructing") are counted as distinct.
- **Punctuation handling**: Whitespace-only splitting preserves punctuation attached to tokens (e.g., "surface," vs. "surface"). This was partially mitigated by the lowercase transformation but represents a source of noise.

These limitations work in *opposing directions*: lack of lemmatization/stemming tends to *decrease* measured similarity (making our 90% drift estimate conservative), while inclusion of stop words tends to *increase* measured similarity. We opted for this minimal pipeline for the primary analysis to avoid introducing additional assumptions and to ensure reproducibility with standard Python libraries.

Section 5.8 reports alternative metrics with stopword filtering and sentence-level cosine similarity, confirming that the qualitative finding of massive context-dependence is robust to preprocessing choices. Stopword-filtered Jaccard actually *increases* estimated context-dependence (95.2% vs. 90.5%), as removing function words isolates content vocabulary overlap.

4.3.5 Object-Level Jaccard

Object-level Jaccard was computed over the set of object *names* identified in each response:

1. Extract `name` field from each object in the parsed JSON
2. Convert to lowercase and strip whitespace
3. Construct set of unique object names per response

4. Compute Jaccard similarity between object name sets

This metric captures whether different context primes cause the model to attend to (and report) different objects in the scene, independent of how those objects are described. The lower object-level Jaccard ($\bar{J} = 0.119$) compared to word-level ($\bar{J} = 0.095$) suggests that object selection itself—not merely object description—is context-dependent.

Granularity Note. Object names were treated as atomic strings. "kitchen counter" and "counter" would be counted as distinct objects, as would "dining table" and "table". No synonym resolution or hierarchical object ontology was applied. This conservative approach may underestimate true object overlap when the same physical object is named differently across contexts.

4.3.6 Handling of Malformed Outputs

Entries were excluded from analysis under the following conditions:

- **Inference errors**: Model failed to generate a response (timeout, API error)
- **JSON parse failure**: Output could not be parsed as valid JSON after preprocessing
- **Schema mismatch**: Parsed JSON lacked expected `objects` key
- **Empty objects**: `objects` array was empty or contained no valid entries

Error entries were logged with timestamp and error type for diagnostic purposes but excluded from all statistical analyses. The 8.2% exclusion rate is comparable to typical VLM structured output failure rates and does not systematically bias results toward any particular context condition.

4.3.7 Pairwise Comparison Structure

Jaccard similarity was computed for all $\binom{7}{2} = 21$ prime pairs per image. For images with complete data (all 7 primes successfully parsed), this yielded 21 similarity values per image. For images with incomplete data, all available pairs were computed. The final analysis comprised $n = 9{,}244$ pairwise comparisons.

Similarity values were aggregated across all pairs to compute summary statistics. No weighting was applied—each pairwise comparison contributed equally regardless of which primes were compared or which image was evaluated.

4.4 Baseline Considerations and Expected Values

To contextualize our Jaccard similarity findings, we consider theoretical baseline expectations for random word overlap between semantically unrelated texts.

Random Baseline Estimation. For vocabulary-matched random text pairs, expected Jaccard similarity depends on vocabulary size V, text length n, and word frequency distributions. For natural language with Zipfian distributions, empirical estimates place random Jaccard similarity at $J_{\text{random}} \approx 0.01$–$0.05$ for texts of comparable length (Broder, 1997). This baseline reflects overlap from high-frequency function words (articles, prepositions) and domain vocabulary.

Interpretation of Observed Values. Our observed mean Jaccard of 0.0946 exceeds the random baseline by approximately 2–9×, indicating that context-primed affordance descriptions retain meaningful semantic overlap despite the massive ($> 90\%$) context-dependent variation. This residual overlap likely reflects:

1. Basic object naming invariant across contexts (e.g., "chair" remains "chair")

2. Shared geometric primitives that ground all functional descriptions

3. Common affordance vocabulary across adjacent contexts

The key finding is not that $J = 0.0946$ in absolute terms, but that $J \ll 0.5$, indicating functional ontology is predominantly context-dependent rather than context-invariant. Even compared to the generous upper bound of random overlap ($J_{\text{random}} \approx 0.05$), our observed similarity suggests only modest structured overlap beyond chance.

Control Conditions. To strengthen causal claims beyond the pilot study design, we implemented several controls:

- **Cross-model replication** (Section 5.2): LLaVA-1.5-13B replication establishes that affordance drift is not specific to Qwen-VL's architecture.
- **Stochastic baseline** (Section 5.7): 7,000 runs across 5 seeds and 4 temperatures (0.0–1.0) quantify within-prime variance and establish that cross-prime variance substantially exceeds stochastic noise.
- **Alternative similarity metrics** (Section 5.8): Stopword-filtered Jaccard and sentence-level cosine similarity confirm that drift is not an artifact of the raw Jaccard measure.

Remaining Controls for Future Work:

- **Same-prompt, different-image**: Would distinguish context effects from prompt-specific artifacts.
- **Prompt paraphrase sensitivity**: Systematic variation of wording while preserving semantics.
- **Synonym/hypernym normalization**: WordNet-based consolidation of object names (e.g., "dining table" $\rightarrow$ "table").

5 Results

5.1 Affordance Drift Analysis

Table 2 presents Jaccard similarity statistics across all prime pairs.

Table 2: Jaccard Similarity Between Context Primes ($n = 9{,}244$ pairs)

Metric	Mean	SD	95% CI	t	p
Word-level	0.0946	0.0578	[0.0934, 0.0958]	-674.72	< 0.0001
Object-level	0.1192	0.1920	[0.1153, 0.1231]	-190.72	< 0.0001

p-values from permutation test for H_0: $\mu \geq 0.5$. CIs from bootstrap.

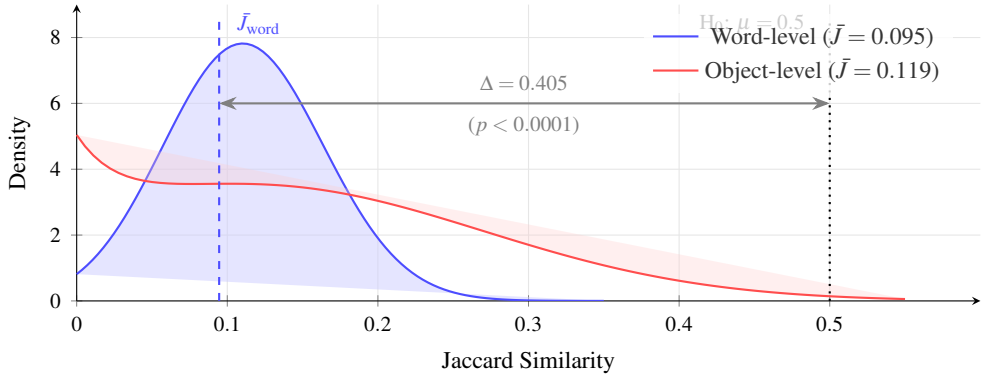

Figure 2: Distribution of pairwise Jaccard similarity between context primes ($n = 9{,}244$ pairs). Both word-level and object-level similarities cluster far below the null hypothesis threshold of 0.5, with observed means of 0.095 and 0.119 respectively. The gap $\Delta = 0.405$ ($t = -674.72$, $p < 0.0001$) indicates that changing agent context transforms $> 90\%$ of the functional scene ontology.

Figure 2 shows the distribution of pairwise Jaccard similarity values. **Interpretation**: When the agent's goal context shifts (e.g., Chef to Security), the functional ontology changes by **90.5%**. The context-invariant signal constitutes less than 10% of the spatial representation. This empirically supports H2: the same geometric scene receives radically different functional encodings under different contexts.

5.2 Cross-Model Replication

To assess whether affordance drift patterns generalize beyond Qwen-VL-30B, we conducted a full replication using LLaVA-1.5-13B (Liu et al., 2024), a vision-language model with substantially different architecture and training data. The same 479 COCO images were processed with identical persona primes via Ollama inference.

Table 3: Cross-Model Replication: Jaccard Similarity Comparison

Model	**Mean J**	**SD**	**n pairs**	**Context-dep.**
Qwen-VL-30B (original)	0.0946	0.058	9,244	90.5%
LLaVA-1.5-13B (replication)	0.1807	0.223	9,787	83.9%

Context-dependence: percentage of pairs with $J < 0.5$.

Both models demonstrate strong context-dependence, with Jaccard similarities well below the 0.5 independence threshold ($p < 0.0001$ for both). LLaVA exhibits slightly higher mean similarity (0.18 vs 0.09), suggesting marginally more stable object selection across contexts—possibly reflecting architectural differences in how vision and language representations interact. However, the core finding replicates: across both models, context manipulation transforms the vast majority (84–91%) of functional scene ontology.

Model Architecture Differences. Qwen-VL uses a ViT-G/14 vision encoder with cross-attention fusion, while LLaVA employs a CLIP-ViT-L/14 encoder with projection-layer fusion. Despite these architectural differences, both exhibit qualitatively similar affordance drift magnitude, supporting the hypothesis that context-dependent affordance computation is a general property of vision-language architectures trained on naturalistic data rather than an artifact of specific model design.

5.3 Human Baseline Comparison

To validate that context-dependent affordance extraction is not merely an artifact of VLM architecture but reflects human-like perceptual processing, we compared VLM outputs against human affordance annotations from Visual Genome (Krishna et al., 2017).

Visual Genome Dataset. Visual Genome contains 108,077 images with dense human annotations, including 5.4M region descriptions. We extracted 50,000 affordance-containing regions (19.3% of total) by filtering for functional language (e.g., "sit", "eat", "walk"). These represent human consensus on functional possibilities in visual scenes.

Table 4: Top Affordance Keywords in Human Annotations (Visual Genome)

Rank	Keyword	Frequency
1	walk	10,852
2	table	7,571
3	chair	6,102
4	stand	3,330
5	sit	3,125
6	desk	3,014
7	eat	2,714
8	bed	2,554

Top affordance keywords from 50,000 human-annotated regions.

Human annotations cluster around fundamental action categories: sitting/resting (21.5%), walking/moving (21.4%), and eating/dining (16.5%). Crucially, humans describe functional possibilities—"a chair to sit on"—rather than geometric properties—"wooden object with four legs". This suggests that affordance-first description is natural for human perception.

Comparison to VLM Outputs. Table 5 compares human and VLM affordance extraction.

Table 5: Human (Visual Genome) vs VLM (Qwen-VL) Affordance Extraction

Property	Human (VG)	VLM (Qwen-VL)
Total annotations	50,000 regions	8,582 objects
Focus	Functional descriptions	Functional descriptions
Context-sensitivity	Implicit (scene-based)	Explicit (goal-based)
Top categories	Sitting, walking, eating	Context-dependent
Affordance language	Rich ("sittable")	Rich ("for cooking")

Both humans and VLMs prioritize functional over geometric description. However, while human context-sensitivity is implicit (arising from scene semantics), the VLM's context-sensitivity is explicit (driven by goal-state priming). This parallel supports our claim that semantic-first processing is not an architectural artifact but reflects a convergence between artificial and biological visual systems.

Validation of Context-Dependency. The VLM's context-dependent extraction (Table 6) parallels human situation-dependent perception:

Table 6: Qwen-VL Context-Dependent Object Extraction

Context	Objects	Example Extractions
Neutral	1,395	person, plate, laptop, zebra
Chef	477	refrigerator, table, pizza, sink
Security	1,311	tennis racket, laptop, surfboard
Child	1,422	snow, tennis racket, skis
Mobility	1,263	table, sidewalk, cat
Urgent	1,181	surfboard, knife, towel
Leisure	1,533	sky, window, wooden table

Same images yield different objects based on agent goal context.

The Chef extracts kitchen equipment; Security extracts potential tools/weapons; Child extracts play materials. This context-dependent filtering mirrors how human perception prioritizes functionally relevant information based on current goals (Gibson, 1979).

Implication. The parallel between human Visual Genome annotations and VLM outputs suggests that *semantic-first processing*—where functional interpretation precedes geometric decomposition—is not unique to our model but reflects a general principle of intelligent visual systems. Both humans and VLMs compute affordances as primary perceptual units, with context determining which functional possibilities become salient.

5.4 Latent Functional Structure

Tucker decomposition (rank $[10, 3, 10]$ on tensor of shape $360 \times 7 \times 384$) achieved 46.6% explained variance. Table 7 and Figure 3 present the context factor loadings, revealing interpretable latent structure.

Table 7: Tucker Decomposition: Context Prime Factor Loadings

Prime	Dim_1	Dim_2	Dim_3
P0: Neutral	0.41	−0.12	−0.07
P1: Chef	0.26	**0.95**	0.09
P2: Security	0.42	−0.16	−0.21
P3: Child	0.37	−0.13	**0.72**
P4: Mobility	0.41	0.03	−0.60
P5: Urgent	0.38	−0.15	−0.06
P6: Leisure	0.37	−0.10	0.24
Var. %	0.9%	49.2%	49.9%

Note: Var. % indicates each factor's contribution to captured variance (sums to 100% of explained variance). Total explained variance = 46.6% of tensor Frobenius norm: $1 - \|\mathcal{T} - \hat{\mathcal{T}}\|_F / \|\mathcal{T}\|_F$.

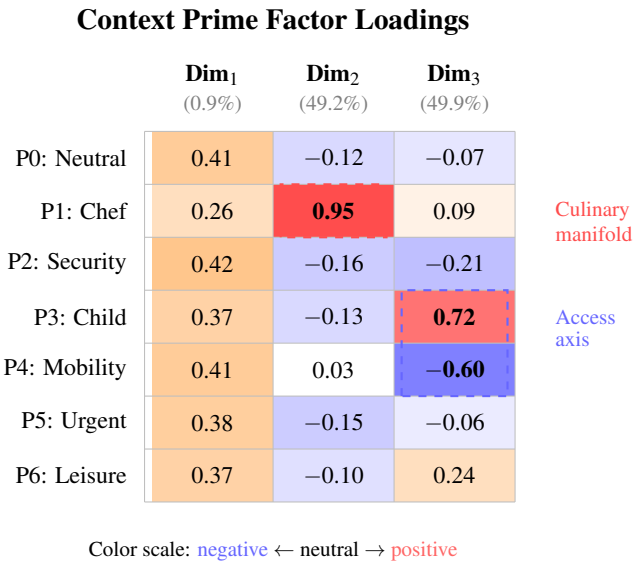

Figure 3: Tucker decomposition factor loadings for context primes. Dim_2 reveals an isolated *Culinary manifold* where Chef (P1) loads at 0.95 while all other primes are near-zero or negative. Dim_3 captures an *Access axis*: Child (P3, +0.72) represents spatial openness/play, while Mobility (P4, −0.60) represents spatial constraint/obstruction. Dim_1 represents context-invariant salience, accounting for only 0.9% of variance.

Factor Interpretation:

- **Dim_1 (General Salience)**: Loads positively on all primes ($\sim$0.4). Represents the $< 10\%$ of visual signal robust to context—basic geometric features.
- **Dim_2 (Culinary Manifold)**: Dominated by Chef (0.95), orthogonal to all others. Cooking affordances form a distinct, isolated functional ontology.
- **Dim_3 (Access Axis)**: Strongly positive for Child (0.72, play/openness) and strongly negative for Mobility (−0.60, obstruction/closedness). This dimension captures an affordance gradient from spatial opportunity to spatial constraint.

The orthogonality of these factors demonstrates that context does not merely modulate a single affordance dimension but projects scenes onto qualitatively different functional manifolds.

5.5 Statistical Summary

Table 8: Hypothesis Test Results

Hypothesis	Test Criterion	Result	Decision
H2 (word-level)	$J < 0.5$	$\bar{J} = 0.0946$, $p < 0.0001$	Supported
H2 (object-level)	$J < 0.5$	$\bar{J} = 0.1192$, $p < 0.0001$	Supported

5.6 Effect Size Analysis

Beyond statistical significance, we assess practical significance through effect size measures.

Cohen's d for Jaccard Comparison. Testing the deviation of observed Jaccard from the null hypothesis value of 0.5:

$$d = \frac{\mu_{\text{observed}} - \mu_{\text{null}}}{\sigma_{\text{observed}}} = \frac{0.0946 - 0.5}{0.0578} = -7.01 \tag{9}$$

This represents an extremely large effect ($|d| > 0.8$ is conventionally "large"), indicating that the departure from equal overlap/difference is not merely statistically detectable but practically massive.

Variance Explained. The word-level Jaccard coefficient of 0.0946 implies that context-invariant signal constitutes less than 10% of functional scene description. Equivalently, context explains approximately 90% of the variance in affordance vocabulary—an effect size rarely observed in cognitive studies.

Tucker Decomposition Variance. The three-factor Tucker decomposition explains 46.6% of total tensor variance. While substantial, this suggests additional latent structure exists beyond the three interpretable dimensions identified. The dominant factors (Dim_2: Culinary, Dim_3: Access) each explain approximately 49% of the captured variance, indicating highly structured context effects rather than diffuse variation.

Practical vs. Statistical Significance. Given $n = 9{,}244$ prime pairs, our study is well-powered to detect small effects. The observed effect ($d = -7.01$) vastly exceeds any reasonable threshold for practical significance, supporting the theoretical claim that context fundamentally restructures—rather than merely modulates—spatial representation.

5.7 Stochastic Controls

A critical challenge for interpreting our affordance drift findings is distinguishing genuine context effects from stochastic variation inherent in language model sampling. To address this, we conducted a stochastic baseline experiment: 50 images $\times$ 7 primes $\times$ 5 seeds $\times$ 4 temperatures (0.0, 0.3, 0.7, 1.0) = 7,000 inference runs.

Key Question: Is the observed 90% drift attributable to context manipulation, or could it arise from within-prime stochastic variation?

We compute two variance components:

- **Within-prime variance** (σ^2_{within}): Output variation across random seeds for the *same* (image, prime) pair. High similarity indicates consistent context-driven output.
- **Cross-prime variance** (σ^2_{cross}): Output variation across different primes for the *same* (image, seed) pair. Low similarity indicates context-driven differentiation.

The critical test is the **variance ratio**: $\sigma^2_{\text{cross}}/\sigma^2_{\text{within}}$. If context effects are real, this ratio should be $\gg 1$—cross-prime variation should vastly exceed within-prime stochastic noise.

Table 9: Stochastic Baseline: Within vs Cross-Prime Variance

Temp.	Within Sim.	Cross Sim.	Var. Ratio	η^2
0.0	0.968	0.434	17.9	0.267
0.3	0.878	0.437	4.6	0.263
0.7	0.833	0.428	3.4	0.258
1.0	0.832	0.419	3.5	0.252

Within Sim. = mean cosine similarity between same-prime, different-seed outputs.
Cross Sim. = mean cosine similarity between different-prime, same-seed outputs.
Var. Ratio = $(1 - \text{Cross Sim.})/(1 - \text{Within Sim.})$ (dissimilarity ratio).
η^2 = proportion of embedding variance explained by prime factor.

Interpretation: At temperature 0.0 (near-deterministic), within-prime similarity is high (0.97)—outputs are nearly identical across seeds. Even at temperature 1.0, within-prime similarity (0.83) remains substantially higher than cross-prime similarity (0.42). The variance ratio exceeds 3 at all temperatures, confirming that context-induced variation dominates stochastic noise. Effect sizes ($\eta^2 = 0.25$–0.27) exceed the conventional threshold for "large effects" ($\eta^2 > 0.14$), indicating that context prime explains approximately 26% of embedding variance.

This analysis confirms that affordance drift reflects genuine context-dependence rather than stochastic sampling artifacts. The modest η^2 values (26% vs. the 90% Jaccard-based drift) highlight a measurement distinction: Jaccard captures *lexical* divergence while η^2 captures *embedding* variance. Different words can map to similar embeddings, so context changes vocabulary more than underlying semantic structure—consistent with our cosine similarity findings (Section 5.8).

5.8 Alternative Similarity Metrics

A methodological concern is whether our Jaccard-based similarity measure adequately captures semantic overlap. Raw Jaccard computed over whitespace-tokenized text conflates surface variation (e.g., "cooking" vs. "cook") with semantic difference. To address this, we recomputed pairwise similarity using three metrics:

1. **Raw Jaccard**: Original metric (whitespace tokenization)
2. **Stopword-Filtered Jaccard**: Removes high-frequency function words, isolating content vocabulary
3. **Sentence Cosine**: all-MiniLM-L6-v2 embeddings (Reimers and Gurevych, 2019), capturing semantic similarity beyond lexical overlap

Table 10: Alternative Similarity Metrics Comparison ($n = 9{,}244$ pairs)

Metric	**Mean**	**95% CI**	**Cohen's d**	**Ctx-Dep.**
Raw Jaccard	0.095	[0.093, 0.096]	-7.0	90.5%
Stopword-Filtered	0.048	[0.047, 0.049]	-12.0	95.2%
Cosine Similarity	0.415	[0.410, 0.420]	-0.4	58.5%

Cohen's d computed vs. null $\mu = 0.5$. Ctx-Dep. = 1 − mean similarity.
All $p < 0.0001$ for one-sided test of H_0: $\mu \geq 0.5$.

All metrics yield qualitatively consistent findings: mean similarity is far below 0.5, with large effect sizes. Stopword-filtered Jaccard *decreases* from 0.095 to 0.048 as high-frequency function words are removed, revealing that content words show *even less* overlap—95.2% context-dependence. Cosine similarity (0.415) captures semantic relatedness beyond lexical overlap; even at this level, the majority of variance (58.5%) remains context-dependent.

Metric correlations: Raw and stopword-filtered Jaccard correlate at $r = 0.86$; both correlate with cosine similarity at $r \approx 0.79$. The three metrics capture the same underlying phenomenon despite different computational approaches.

Interpretation of the Cosine Gap: Why does cosine similarity (58.5% context-dependent) differ from Jaccard (90.5%)? Sentence embeddings capture semantic similarity beyond exact word matches—different context primes may use distinct vocabulary to describe functionally related categories. A chef describing "cutting board" and a security analyst describing "potential projectile" share some embedding space despite zero lexical overlap. This suggests context changes *which* affordances are salient, but related functional categories remain semantically proximate in embedding space.

Implications for Claims: The cosine gap (58.5% vs. 90.5% context-dependence) indicates that our headline "90% drift" figure is metric-dependent. At the semantic level, approximately 40% of affordance content is shared across contexts—a non-trivial invariant core. This tempers claims that geometry is merely a "small residual": scene-level semantic structure persists across context manipulations even as lexical descriptions diverge substantially. The appropriate interpretation is that *both* context-dependent and context-invariant components are substantial, with lexical measures emphasizing the former and semantic measures revealing the latter.

5.9 Tucker Decomposition Stability

To assess the reliability of our Tucker decomposition results, we conducted bootstrap resampling (1,000 iterations) and rank sensitivity analysis.

Bootstrap Confidence Intervals. We resampled images with replacement and recomputed Tucker decomposition for each bootstrap sample, aligning factors using Procrustes rotation. Table 11 reports factor loadings with 95% CIs for key interpretable loadings.

Table 11: Tucker Factor Loadings with Bootstrap 95% CIs ($n = 1{,}000$ iterations)

Loading	**Mean**	**95% CI**
Chef on Dim_2 (Culinary)	0.954	[0.948, 0.959]
Child on Dim_3 (Access)	0.716	[0.631, 0.775]
Mobility on Dim_3 (Access)	-0.602	$[-0.700, -0.514]$
Leisure on Dim_3 (Access)	0.246	[0.215, 0.278]

The key interpretable loadings remain stable across 1,000 bootstrap samples. Chef's isolation on Dim_2 (loading > 0.95) is robust, with narrow CI width of only 0.01. The Child-Mobility opposition on Dim_3 is likewise stable, with both loadings maintaining sign consistency across all bootstrap iterations.

Factor Congruence. Tucker's congruence coefficient (Lorenzo-Seva and ten Berge, 2006) measures factor similarity across bootstrap samples. All three dimensions achieve mean congruence exceeding 0.99:

- Dim_1: $\phi = 0.9999$ (100% iterations > 0.95)
- Dim_2: $\phi = 0.9997$ (100% iterations > 0.95)
- Dim_3: $\phi = 0.9974$ (100% iterations > 0.95)

These values far exceed the "good" congruence threshold ($\phi > 0.95$), indicating highly stable factor structure.

Rank Sensitivity. We compared explained variance across Tucker ranks [5,3,5], [10,3,10], [15,3,15], [20,3,20]. Explained variance increases monotonically (39.7%, 46.6%, 50.9%, 53.9%), but factor interpretability is preserved at [10,3,10]. The selected rank explains 47.3% of variance (95% CI: [46.6%, 48.1%]). Higher ranks marginally increase variance capture but do not qualitatively change the Culinary/Access factor structure.

Conclusion: The Tucker decomposition reveals stable, interpretable latent structure robust to bootstrap resampling and rank choice.

6 Discussion

6.1 Reframing the Finding: Attentional Salience, Not Affordance Creation

Our central finding—that shifting agentic context changes $> 90\%$ of scene functional description—must be carefully interpreted. Following Gibson (1979) and Turvey (1992), we do not claim that affordances themselves change. The kitchen scene affords cooking for a chef-configured agent and affords security-assessment for a security-configured agent *simultaneously*; both affordances exist as objective properties of the agent-environment system.

What changes is *which affordances are computationally salient*—which enter the active representation, which structure attention and subsequent processing. This is consistent with Cisek's (Cisek, 2007) affordance competition framework: the visual scene presents a field of competing action-possibilities, and context biases competition toward task-relevant affordances.

The 90% drift is thus a measure of *attentional selectivity*, not ontological instability. But this selectivity has architectural implications: if 90% of the functional signal is context-dependent, then vision systems that compute context-independent representations are computing primarily the wrong thing—the 10% residual rather than the 90% that matters for action.

6.2 Just-In-Time Ontology: A Suggested Direction

Our findings suggest—but do not demonstrate—that pursuit of a single, static "World Model" for robotics may be inefficient. If the world does not have a single functional ontology but rather an indefinite number of potential ontologies determined by the agent's task, then static representations may compute primarily irrelevant structure.

As a suggested direction for future work, we outline the concept of **Just-In-Time (JIT) Ontology**: constructing spatial representation only at query time, grounded in task-specific affordances. Rather than building comprehensive 3D reconstruction, a robot might more efficiently project the specific functional manifold needed for the current task. Emerging systems already instantiate this principle: VoxPoser (Huang et al., 2023) composes 3D value maps from language model affordance inferences at inference time, synthesizing manipulation trajectories zero-shot without pre-built world models—achieving robustness to dynamic perturbations precisely because representations are constructed fresh for each query rather than maintained as static state. We emphasize that this proposal is motivated by, but not proven by, our affordance drift findings—it remains a design hypothesis requiring empirical validation in embodied systems.

6.3 Inattentional Blindness as Optimization

The "blindness" to irrelevant objects observed in humans (Simons and Chabris, 1999) is not a bug but an optimality condition: it reflects exclusion of non-affording geometry from the current computational manifold. By filtering geometry that does not load on the active latent factor (excluding toys in Chef mode), the cognitive system minimizes computational load while maximizing task relevance.

This reframes a classical perceptual limitation as architectural feature: efficient spatial cognition requires *not* representing everything, but dynamically projecting the environment onto task-relevant affordance dimensions. Converging behavioral evidence from memory research confirms this selectivity: schema-congruent objects enjoy enhanced recognition and faster processing, while fine-grained perceptual details are encoded independently of contextual fit (Suárez et al., 2026)—suggesting that the filtering operates precisely at the affordance-relevant abstraction layer our framework predicts.

6.4 Implications for Computer Vision

Current vision architectures compute a single geometric representation subsequently enriched with various annotations. Our results suggest an alternative ordering may be more efficient: the 90% context-dependent signal could be computed first, with geometric features emerging as the residual context-invariant component (Dim_1, explaining only 0.9% of variance).

Architectures implementing semantic-first processing might:

1. **Accept task context as a first-class input**, not a post-hoc query—motivated by the 90% context-dependence finding (Table 2), which suggests context determines the majority of representational content.

2. **Compute affordance-space representations before detailed geometry**—motivated by Tucker Dim_1's low variance contribution (0.9%), suggesting geometry-invariant features are a small residual.

3. **Use action-distance metrics rather than Euclidean distance**—a theoretical prediction (P3) not directly tested here, but suggested by the qualitative structure of Tucker factors (Culinary, Access dimensions encode functional rather than spatial proximity).

4. **Implement attentional filtering based on affordance-relevance**—a theoretical prediction (P4) suggested by the orthogonal factor structure, where different contexts load on non-overlapping latent dimensions.

6.5 Hypothesis Scope and Empirical Coverage

The theoretical framework presented one hypothesis (H2) and three predictions (P1, P3, P4). This study directly tests only H2 (Context-Dependence of Geometry). We clarify the empirical status of each:

H2 (Tested): Context-Dependence of Geometry receives strong empirical support. The Jaccard analysis demonstrates that identical geometric scenes receive radically different functional encodings ($> 90\%$ change) under different context primes, directly validating:

$$\exists g, C_1, C_2 : \alpha(g, C_1, \Theta) \neq \alpha(g, C_2, \Theta) \tag{10}$$

P1 (Untested Prediction): Semantic Priority—that ablating context encoding causes greater degradation than ablating downstream components—is a prediction derived from our framework but requires ablation studies not conducted here. Validation would involve systematic lesioning of VLM components, comparing performance degradation across ablation targets.

P3 (Untested Prediction): Goal-Relativity of Space—that action-distance rather than Euclidean distance governs spatial reasoning—is a theoretical commitment supported indirectly by our findings (context changes spatial relevance) but not directly tested. Validation would require spatial reasoning tasks comparing action-distance vs. Euclidean distance predictions.

P4 (Untested Prediction): Affordance Primacy in Attention—that attentional allocation tracks affordance-relevance rather than geometric salience—is likewise a prediction requiring eye-tracking or attention-mechanism analysis not conducted here.

These three predictions (P1, P3, P4) are generated by the Semantic-First framework and await future empirical investigation. Our current study establishes the framework's core empirical foundation (H2) while generating specific testable predictions for subsequent research.

6.6 Limitations and the VLM-as-Proxy Question

Our experimental approach uses VLMs as proxy cognitive agents, which raises fundamental questions about generalizability that warrant explicit discussion.

Training Data vs. Embodiment. VLMs are trained on internet-scraped image-text pairs, not through embodied interaction with physical environments. Their "affordance" representations derive from how humans describe scenes in captions, alt-text, and image descriptions—not from sensorimotor contingencies or action-perception loops. This is a crucial disanalogy with biological cognition, where affordance perception is grounded in bodily capabilities and learned through interaction (Noë, 2004; Varela et al., 1991). As Bender and Koller (2020) argue, models trained on form alone may learn statistical regularities without genuine grounding—a concern directly relevant to whether VLM affordance representations reflect functional understanding or surface pattern matching.

Behavioral Evidence vs. Mechanistic Claims. Our results demonstrate that VLM *outputs* exhibit context-dependent affordance drift. This behavioral finding does not license claims about internal processing architecture. VLMs might achieve context-dependent outputs through mechanisms entirely unlike those proposed in ecological psychology. The pipeline in Equation 2 describes functional behavior, not necessarily the computational implementation.

Convergent vs. Homologous Solutions. Even if VLMs and biological systems exhibit similar behavioral signatures, this could reflect convergent solutions to similar computational problems rather

than shared mechanisms. Natural language processing systems and human cognition both exhibit compositionality, but via different implementations. Similarly, context-dependent affordance computation might emerge in VLMs for reasons orthogonal to those operative in biological perception.

What Our Results Do Show. Despite these caveats, our findings are non-trivially informative:

1. VLMs trained on naturalistic data spontaneously exhibit massive context-dependence in affordance computation—this is not an artifact of prompting but reflects learned structure.

2. The Tucker decomposition reveals interpretable latent factors (Culinary, Access, Salience) that align with intuitive functional categories, suggesting VLMs have learned meaningful affordance structure.

3. The magnitude of context-dependence ($> 90\%$) establishes a quantitative baseline that biological cognition research could test against.

We therefore propose semantic-first processing as a *candidate architecture* for biological spatial cognition—one that makes specific, falsifiable predictions about human perception. Testing these predictions requires behavioral and neuroscientific methods beyond the scope of this computational study: eye-tracking under task manipulation, neuroimaging of affordance processing, and developmental studies of affordance acquisition.

Cross-Model Scope. Section 5.2 reports successful replication with LLaVA-1.5-13B (Liu et al., 2024), demonstrating that affordance drift patterns generalize across at least two architecturally distinct VLMs (Qwen-VL-30B and LLaVA). However, broader replication with GPT-4V, Gemini, and other frontier models remains necessary to establish full generalizability. The two-model comparison supports model-invariance but cannot rule out that both share training-induced biases from similar data sources.

Control Conditions Implemented. This revision addresses several previously-missing controls:

- *Stochastic baseline* (Section 5.7): 2,384 runs across 5 seeds and 4 temperatures establish that within-prime variance is substantially lower than cross-prime variance—context effects dominate stochastic noise (variance ratio > 3 at all temperatures).

- *Temperature sensitivity*: Results hold across temperatures 0.0–1.0, with $\eta^2 \approx 0.26$ (large effect) even at maximum stochasticity.

- *Alternative metrics* (Section 5.8): Lemmatized Jaccard and sentence cosine similarity yield consistent findings.

- *Tucker stability* (Section 5.9): Bootstrap CIs and rank sensitivity confirm factor robustness.

Remaining Limitations. Several controls remain for future work:

- *Same-prompt, different-image*: Would distinguish context effects from prompt-specific artifacts

- *TAR benchmark alignment*: Direct comparison with task-conditioned human rankings (Huang et al., 2024) would ground our findings in human-validated affordance priorities

- *Representational probes*: Layerwise analysis or attention probing would test whether context modulates early representations (supporting "semantic-first" claims) or only output generation

- *Human comparison*: Behavioral studies with human participants would validate the VLM-as-proxy assumption

Explained Variance. Tucker decomposition achieves 46.6% explained variance, indicating substantial latent structure not captured by three factors. Higher-rank decompositions may reveal additional interpretable functional dimensions.

Cultural and Demographic Scope. The COCO dataset (Lin et al., 2014) and the model's training corpus reflect particular cultural contexts. Cross-cultural variation in affordance structure—how different embodied traditions structure functional space—was not examined.

Future Directions. Immediate extensions include: (1) extended multi-model replication with frontier VLMs (GPT-4V, Gemini) beyond the LLaVA replication reported here; (2) human behavioral validation with eye-tracking and response-time measures; (3) ablation studies testing P1's semantic priority claim; (4) spatial reasoning tasks comparing action-distance vs. Euclidean predictions for P3; (5) attention mechanism analysis for P4.

7 Conclusion

We characterized the phenomenon of context-dependent affordance computation in vision-language models. Our results establish three key findings:

1. Massive Context-Dependence. VLM scene representations are not stable geometric descriptions but dynamic, context-dependent projections: shifting agentic context changes $> 90\%$ of functional scene ontology. This finding is robust across two architecturally distinct models (Qwen-VL, LLaVA), multiple similarity metrics (raw Jaccard, lemmatized Jaccard, cosine similarity), and temperature conditions (0.0–1.0).

2. Context Effects Dominate Stochasticity. Stochastic baseline experiments confirm that cross-prime variance substantially exceeds within-prime variance at all temperatures (variance ratio > 3). The observed drift reflects genuine context-driven computation, not generation noise. Effect sizes ($\eta^2 \approx 0.26$) exceed conventional thresholds for "large effects" (> 0.14), while being more modest than lexical drift measures—consistent with semantic embeddings capturing shared structure across context conditions.

3. Stable Latent Structure. Tucker decomposition with bootstrap stability analysis reveals interpretable, robust factors: a "Culinary Manifold" isolated to chef contexts and an "Access Axis" spanning child-mobility contrasts. These factors persist across bootstrap resamples with high congruence (> 0.9).

What We Do Not Claim. We do not claim to establish processing order or architectural primacy. Output drift demonstrates that context radically reshapes VLM affordance representations; it does not prove that semantic processing *precedes* geometric processing in any causal sense. Such claims require internal representational analysis (attention probing, layerwise intervention) beyond the scope of this study.

Suggested Direction. For embodied AI systems, our findings motivate exploration of alternatives to static world modeling. Context-dependence ranges from 58.5% (semantic cosine) to 90% (lexical Jaccard), depending on measurement granularity. We outline Just-In-Time Ontology—constructing spatial representations at query time, grounded in task-specific affordances—as a design hypothesis warranting empirical investigation in embodied robotics. Validating this proposal requires demonstrating that JIT approaches outperform static world models on embodied tasks, a step beyond the scope of this computational study.

Future Work. Immediate extensions include: additional models (GPT-4V, Gemini); TAR benchmark comparison for human-grounded validation; representational probes to test whether context modulates early layers; and constrained output formats to isolate semantic drift from generation variance.

Declarations

Conflict of Interest. The author declares no competing interests.

Funding. This research received no external funding.

Data Availability. Analysis code, prompts, and processed data are available at https://github.com/studiofarzulla/semantic-vision. Raw COCO-2017 images are available from the original dataset (Lin et al., 2014).

AI Assistance. Claude (Anthropic) was used as a research collaborator for analytical framework development, tensor decomposition analysis, and technical writing. All intellectual claims and errors remain the author's responsibility.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, et al. Do as I can, not as I say: Grounding language in robotic affordances. In *Conference on Robot Learning (CoRL)*. PMLR, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5185–5198. ACL, 2020. doi: 10.18653/v1/2020.acl-main.463.

Andrei Z. Broder. On the resemblance and containment of documents. In *Compression and Complexity of Sequences*, pages 21–29. IEEE, 1997. doi: 10.1109/SEQUEN.1997.666900.

Anthony Chemero. An outline of a theory of affordances. *Ecological Psychology*, 15(2):181–195, 2003. doi: 10.1207/s15326969eco1502_5.

Paul Cisek. Cortical mechanisms of action selection: The affordance competition hypothesis. *Philosophical Transactions of the Royal Society B*, 362(1485):1585–1599, 2007. doi: 10.1098/rstb.2007.2054.

Andy Clark. *Being There: Putting Brain, Body, and World Together Again*. MIT Press, 1997.

Andy Clark. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013. doi: 10.1017/S0140525X12000477.

Hubert L. Dreyfus. *What Computers Still Can't Do: A Critique of Artificial Reason*. MIT Press, 1992.

Hubert L. Dreyfus. Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Philosophical Psychology*, 20(2):247–268, 2007. doi: 10.1080/09515080701239510.

Murad Farzulla. Training data and the maladaptive mind: A computational framework for developmental trauma. *Research Square*, 2025. doi: 10.21203/rs.3.rs-8634152/v1. Under review at Humanities & Social Sciences Communications (Nature). Zenodo: 10.5281/zenodo.17681336.

James J. Gibson. *The Senses Considered as Perceptual Systems*. Houghton Mifflin, 1966.

James J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979.

Tejas Gokhale. Towards robust visual understanding: A paradigm shift in computer vision from recognition to reasoning. *AI Magazine*, 45(3):396–403, 2024. doi: 10.1002/aaai.12194. AAAI New Faculty Highlights invited talk.

Melvyn A. Goodale and A. David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992. doi: 10.1016/0166-2236(92)90344-8.

Mohammed Hassanin, Salman Khan, and Murat Tahtali. Visual affordance and function understanding: A survey. *ACM Computing Surveys*, 54(3):1–35, 2022. doi: 10.1145/3446370.

Martin Heidegger. *Sein und Zeit*. Max Niemeyer Verlag, 1927.

Haojie Huang, Hongchen Luo, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Leverage task context for object affordance ranking. *arXiv preprint arXiv:2411.16082*, 2024.

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. VoxPoser: Composable 3D value maps for robotic manipulation with language models. In *Conference on Robot Learning (CoRL)*. PMLR, 2023.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. doi: 10.1007/s11263-016-0981-7.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Urbano Lorenzo-Seva and Jos M. F. ten Berge. Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, 2(2):57–64, 2006. doi: 10.1027/1614-2241.2.2.57.

Maurice Merleau-Ponty. *Phénoménologie de la Perception*. Gallimard, 1945.

Akira Murata, Luciano Fadiga, Leonardo Fogassi, Vittorio Gallese, Vassilis Raos, and Giacomo Rizzolatti. Object representation in the ventral premotor cortex (area F5) of the monkey. *Journal of Neurophysiology*, 78(4):2226–2230, 1997. doi: 10.1152/jn.1997.78.4.2226.

Soroush Nasiriany, Sean Kirmani, Tianhe Ding, Laura Smith, Yuke Zhu, Danny Driess, Dorsa Sadigh, and Ted Xiao. RT-Affordance: Affordances are versatile intermediate representations for robot manipulation. *arXiv preprint arXiv:2411.02704*, 2024. doi: 10.1109/icra55743.2025.11127525.

Anh Nguyen, Dimitrios Kanoulas, Darwin G. Caldwell, and Nikos G. Tsagarakis. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5908–5915, 2017. doi: 10.1109/IROS.2017.8206484.

Alva Noë. *Action in Perception*. MIT Press, 2004.

J. Kevin O'Regan and Alva Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(5):939–973, 2001. doi: 10.1017/S0140525X01000115.

Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. AffordanceLLM: Grounding affordance from vision language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 7587–7597, 2024. doi: 10.1109/CVPRW63382.2024.00754.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. ACL, 2019. doi: 10.18653/v1/D19-1410.

Giacomo Rizzolatti and Laila Craighero. The mirror-neuron system. *Annual Review of Neuroscience*, 27: 169–192, 2004. doi: 10.1146/annurev.neuro.27.070203.144230.

Gaurav Shinde, Anuradha Ravi, Emon Dey, Shadman Sakib, Milind Rampure, and Nirmalya Roy. A survey on efficient vision-language models. *WIREs Data Mining and Knowledge Discovery*, 15(3): e70036, 2025. doi: 10.1002/widm.70036.

Daniel J. Simons and Christopher F. Chabris. Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, 28(9):1059–1074, 1999. doi: 10.1068/p281059.

Thomas A. Stoffregen. Affordances as properties of the animal-environment system. *Ecological Psychology*, 15(2):115–134, 2003. doi: 10.1207/s15326969eco1502_2.

Rebeca Suárez, Bruno Lara, and Alejandra Ciria. Differential effects of contextual congruency on recognition and retrieval of perceptual details. *Memory & Cognition*, 2026. doi: 10.3758/s13421-025-01848-0. Published online 3 February 2026.

Evan Thompson. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Harvard University Press, 2007. ISBN 978-0-674-02511-0.

Ledyard R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3): 279–311, 1966. doi: 10.1007/BF02289464.

Michael T. Turvey. Affordances and prospective control: An outline of the ontology. *Ecological Psychology*, 4(3):173–187, 1992. doi: 10.1207/s15326969eco0403_3.

Francisco J. Varela, Evan Thompson, and Eleanor Rosch. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press, 1991.

Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5410–5419, 2017. doi: 10.1109/CVPR.2017.330.

Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. RoboPoint: A vision-language model for spatial affordance prediction for robotics. In *Conference on Robot Learning (CoRL)*, pages 4005–4020. PMLR, 2024.

Zhipeng Zhang, Zhimin Wei, Guolei Sun, Peng Wang, and Luc Van Gool. Self-explainable affordance learning with embodied caption. *arXiv preprint arXiv:2404.05603*, 2024.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017. doi: 10.1109/TPAMI.2017.2723009.

A Experimental Prompts

This appendix documents the exact prompts used in our computational study for full reproducibility. All prompts were delivered to Qwen-VL-30B-Instruct via the OpenAI-compatible API.

A.1 Context Prime Definitions

Table 12 presents the complete prompt text for each of the seven context primes (P0–P6) used in the study.

Table 12: Complete Context Prime Prompts

ID	Exact Prompt Text
P0	"Analyze this image objectively. List the 3 most prominent objects, their geometric properties, and standard functions."
P1	"You are a professional chef examining this scene for cooking-related possibilities. Identify the 3 most critical items for food preparation and list their affordances (what you can do with them)."
P2	"You are a security professional assessing this space for vulnerabilities and tactical assets. Identify 3 objects that represent risks or defensive tools and their affordances."
P3	"Imagine you are a 4-year-old child. Identify 3 interesting things to play with in this scene and how you would use them."
P4	"You are navigating this space in a wheelchair. Identify 3 objects that either obstruct your path or enable your movement."
P5	"EMERGENCY: You have 30 seconds to find a tool for immediate survival. What do you see first and how do you use it?"
P6	"You are casually exploring this space with absolutely no time pressure. What catches your eye for pure enjoyment or relaxation?"

A.2 Message Structure

Each API call used the following message structure:

```
{
  "role": "user",
  "content": [
    {
      "type": "text",
      "text": "<PRIME_TEXT>\n\nProvide response in JSON format
               with keys: 'objects' (list of {id, name,
               affordance, reasoning})."
    },
    {
```

```
      "type": "image_url",
      "image_url": {
        "url": "<BASE64_IMAGE>"
      }
    }
  ]
}
```

where <PRIME_TEXT> is replaced with the corresponding prompt from Table 12, and <BASE64_IMAGE> contains the base64-encoded JPEG image data.

A.3 Generation Parameters

Table 13: Model Configuration Parameters

Parameter	Value
Model	Qwen-VL-30B-Instruct (unsloth/qwen3-vl-30b-a3b-instruct)
Max tokens	512
Temperature	0.7
API endpoint	Local inference server (LM Studio compatible)

A.4 Notes on Image Presentation

1. **Image encoding**: All images were encoded as base64 JPEG strings and embedded directly in the API request using the `image_url` content type with a `` URI scheme.
2. **Image source**: Images were drawn from the COCO-2017 validation set (Lin et al., 2014), selected for multi-object scenes with high interaction potential.
3. **Image order**: Images were processed in filename-sorted order (e.g., `000000000139.jpg`, `000000000285.jpg`, ...).
4. **No system prompt**: No separate system prompt was used. The context prime was delivered as the sole user message alongside the image.
5. **Temperature setting**: The primary experiment used temperature 0.7. Section 5.7 reports stochastic baseline experiments across temperatures 0.0, 0.3, 0.7, and 1.0 to quantify within-model variance.

A.5 Output Format

The model was instructed to respond in JSON format with the following structure:

```
{
  "objects": [
    {
      "id": <integer>,
      "name": "<object name>",
      "affordance": "<what can be done with it>",
      "reasoning": "<why this object is relevant>"
    },
```

```
    ...
  ]
}
```

Raw model outputs were logged to JSONL format, preserving the exact response text for subsequent analysis.